# A Semantic Summarization System: University of Birmingham at TAC 2008

Abdullah Bawakid and Mourad Oussalah

University of Birmingham, School of Engineering
Department of Electronic, Electrical and Computer Engineering
Edgbastoon, Birmingham B15 2TT
{ axb517 , M.oussalah}@bham.ac.uk

## Abstract

Text summarization of document or multi-documents has been acknowledged as one of the most challenging tasks in information system community because of the rich semantic structure of the language and the subjectivity inherent to the summarization task. In this paper, a new query-based extractive summary methodology is put forward. The approach makes use of phrasal decomposition of the text where each sentence is ascribed a scoring function, which will then be used to identify the most relevant sentences in the sequel. The scoring function is expressed as a convex combination of a set of features that are extracted beforehand from the (multi) document(s). Besides, the scoring function includes a semantic similarity evaluation where the WordNet taxonomy is used in conjunction with a variety of other extracted features, as a basis to construct the sentence-sentence semantic similarity. The system architecture as well as its linguistics processing parts are described. Finally, we present the results of our participation in TAC 2008 with possible perspectives.

**Keywords**: Semantic Similarity, WordNet, Linguistic Quantifiers, Text Summarization, Information Retrieval

## 1. Introduction

Text Summarization, as the process of identifying the most salient information in a document or set of documents (for multi-document summarization) and conveying it in less space, became an active field of research in both Information Retrieval (IR) and Natural Language Processing (NLP) communities. Summarization shares some basic techniques with indexing as both are concerned with identification of the essence of a document. Also, high quality summarization requires sophisticated NLP techniques in order to deal with various Parts Of Speech (POS) taxonomy and inherent subjectivity. Typically, one may distinguish various types of summarizers.

Loosely speaking, most common existing summarizers work in an extractive fashion, where portions of the input documents, for instance, sentences, which believed to be more silent, are selected to form the summary. On the other hand, non-extractive does not rely on text selection but rather on a deeper understanding of input text. Query-based summaries are generated in reference to some user query (e.g., summarize a document about an international summit focusing only on the issues related to the environment)

This paper advocates a trade-off methodology between extractive and query-based summarization. The former is due to the fact that the developed methodology uses a scoring function, which uses WordNet taxonomy to generate sentence-sentence semantic similarity as well as a set of extracted features, to quantify the relevance of each sentence. This yields a resulting summary which is nothing else than the most ranked sentences. While the query-based approach is due to the explicit accounting of the topic-sentence semantic similarity in the overall methodology as it will be detailed later on.

The paper describes the system we developed to participate in the update task of TAC 2008. The update summarization task requires participants to submit fluent and organized 100-word multi-document summaries of a set of news articles under the assumption that the user has already read a given set of articles earlier. The summaries to be generated should be relevant to the topic statement given by the user. The purpose of each summary is to inform the reader of new information about a particular topic. The test documents provided were chosen from the AQUAINT-2 collection[1].

The next section gives some background and relates our work with existing summarization systems. In section 3, we give an overview on our system, its main components and how it works. In section 4 we discuss the evaluation performed by NIST on our submitted runs and the obtained results. In section 5 we present some ideas for future work and how the system can be improved.

[1] See http://www.nist.gov/tac/tracks/2008/summarization/ for the detailed task description.

## 2. Background

In the past few years, many multi-document summarization systems have been implemented, most of which are extractive. The key in such systems is to extract the most relevant parts from the source to the user. An example for such systems is MEAD [1] [2] which ranks sentences using a linear combination of features and forms summaries from the highest scoring sentences. MASC [3] is another feature-based summarization system that performs compressions to sentences after the extraction stage.

Our system assigns a score to each sentence in the source documents based on a set of static and dynamic features. Static features include sentences locations and the number of Named Entities (NEs) in each sentence. Dynamic features on the other hand are those that change based on the document sets chosen. The score given for the semantic similarity between a sentence, and the rest of the sentences in the documents set is an example of a dynamic feature employed in our system. Part of our system performs analysis on linguistic quantifiers and combines it with the semantic similarity computing module to form a metric affecting the score given to each sentence.

## 3. System Overview

Figure 1 shows the three main stages involved in generating summaries with our summarizer: *Preprocessing* the source documents, *Extracting and Analyzing* the features, and *Generating the summaries*. The documents are preprocessed first and prepared to extract the features of their sentences.

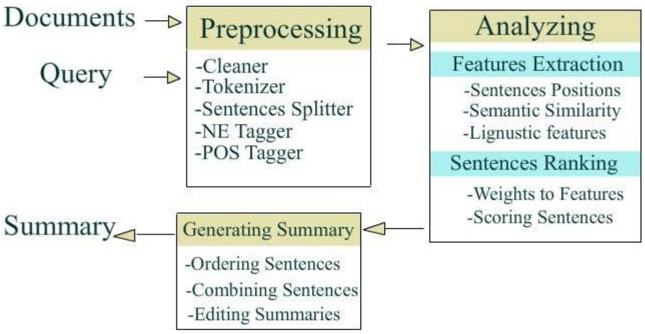

**Figure 1** The Summarizer Architecture.

After extracting the features, a score is computed for each sentence based on the extracted features. The summary is presented at the end by iterating through the sentences and selecting the highest-scoring candidates till the maximum number of words is reached.

The TAC 2008 update task requires participants to submit ~100-word summaries given a group of documents and a topic statement (title and narrative). In our system, the topic statement was treated as the user query. The 100-word limit was met by examining the length of the last sentence appearing in the summary. The 100-word limit was met by iterating through all the highest scoring sentences, starting with the highest rank and proceeding with the next lowest ranked and appending them to the summary until the limit is reached or all candidate sentences are exhausted. If the addition of the last sentence in summary caused the summary length to exceed the limit, it is replaced with the next shorter high scoring sentence. This process can be improved by adding a stage for editing the summary to shorten its length by removing unnecessary information from the summary sentences. Due to time constraints, the editing process was not applied. The following sections examine each of the summarization stages in more details.

### 3.1 Preprocessing

The preprocessing stage involves cleaning the source documents, splitting and annotating the sentences, and extracting the features.

First, unnecessary information and tags are removed from the source documents such as the HTML/XML tags, news agencies names and tables containing numbers. Then, key parts from the documents are extracted such as the publication dates, the documents IDs, and the headlines. The document ID and publication date along with the document name are used to identify each document during the different processing stages. The headline is treated as the document title as explained in the next section. Sentences and word boundaries are then detected and different features are extracted with the help of GATE [4] from the source sentences and the provided user query. The extracted features and annotations include Named Entities in each sentence (Locations, Organizations, and Persons), Part-of-Speech tags (POS), and co reference resolution.

After preprocessing the documents and the queries, the processing stage begins scoring sentences based on the computed/extracted set of features detailed in next section.

### 3.2 Summarization Features

#### 3.2.1 Sentences Location:

The position of sentences in a document can play a significant factor in finding the sentences that are most related to the topic of the document [5]. So, we have decided to take into account the position of sentences when computing the score for each sentence. More weight is given to sentences at the beginning and end of each document than the rest.

### 3.2.2 Named Entities:

Using GATE, it was possible to recognize the Named Entities (NEs) mentioned in each document. The sentences containing more NEs are assumed to be more important than those that contain no NEs. Only the frequency of NEs in each sentence and the document was taken into account when forming the scoring formula.

### 3.2.3 Title / Query

The title of the document, if any, as well as user's query or abstract sentence(s) used to characterize the document or a set of documents are without doubt of paramount importance to quantify the relevance of each sentence/phrase with respect to overall meaning conveyed by the document(s). Therefore, the evaluated semantic similarity of each sentence and title and/or query is explicitly taken into account.

### 3.2.4 Positioning

Typically, the location of the sentence within the document(s) is somehow relevant to context of the document(s) in the sense that sentence located at the beginning or at the end of the document(s) is likely to contain main authors' claim that are developed throughout the whole document(s).

### 3.3 Sentence-sentence semantic Similarity:

To determine the similarity between two sentences, say, a and b, consisting of the sets of terms A and B, each term in A and B is first tagged with their POS (part of speech). It is then determined which noun each adjective describes and which verb each adverb describes. This is done by attempting to find the closest noun or verb following the adjective or adverb, and if none are found the closest noun or adjective preceding the adjective or adverb is used. The adjective and adverb lists are also expanded with exact synonyms from WordNet. The linguistic quantifiers, which may indicate the relative importance of a term within a text, are also associated with nouns in the same way. Typically, linguistic quantifiers are determiners which express information about relative or absolute quantity. The list of linguistic quantifiers is based on Bond's list [12]. In this study, one limited to two classes of linguistic quantifiers: those which induce an increasing order of relevancy like "very", "more", and those inducing a decreasing order like "less", "none", etc.

The similarity score for the sentence is therefore calculated by finding an average of the score for each noun or verb in both of the sentences. First the set of nouns and verbs for each sentence must be found.

$$NVA = \{x \in A : POS(A) = noun \cup verb\}$$

$$NVB = \{x \in B : POS(B) = noun \cup verb\}$$

Next, the best weighted match for each noun and each verb in both sentences must be determined.

$$BA = \{(u,v) : k_{uv} match(u,v) = \max_{i \in NVB}[k_{ui} match(u,i)] \ and\, u \in NVA\}$$

$$BA = \{(u,v) : k_{uv} match(u,v) = \max_{j \in NVA}[k_{jv} match(j,v)] \ and\, v \in NVB\}$$

Where

$$k_{ij} = \begin{cases} 2 \ \ if \ \ both\, i \ and \ \ j \ have\ an\ increase\ quantifier \\ 2 \ \ if \ \ both\, i \ and \ \ j \ have\ an\ decrease\ quantifier \\ 0.5 \ \ if \ \ i \ and \ \ j \ have\ opposit\ quantifier \\ 1 \ \ otherwise \end{cases}$$

And

$$match(u,v) = \begin{cases} (sim(u,v)+1)/2 \quad \text{if matching adjectives or adverbs were found} \\ \text{sim(u,v), otherwise} \end{cases}$$

where sim(u,v) is determined using either Jiang and Conrath's [6] semantic similarity $Sim_{JC}(u,v)$ or Lin's similarity [7] measure $Sim_L(u,v)$.

Finally the semantic similarity between sentence a and b is calculated by average the nouns and verbs semantic similarity as

$$SimSem\,(a,b) = \frac{\sum_{(u,v)\in BA} \mathrm{k}_{\mathrm{uv}}\, match\,(u,v) + \sum_{(u,v)\in BB} \mathrm{k}_{\mathrm{uv}}\, match\,(u,v)}{\sum_{(u,v)\in BA} \mathrm{k}_{\mathrm{uv}} + \sum_{(u,v)\in BB} \mathrm{k}_{\mathrm{uv}}}$$

The effect of this is to get a score which depends on every noun and verb in both sentences. In cases where a matching pair of adjectives or adverbs is found the score will be increased but not exceeding 1. If linguistic quantifiers are found, these are used to weight the average. The word-pair is weighted more highly where matching quantifiers are found and the weighting is reduced when opposite quantifiers are found. The above expression is also used to determine the score attached to the semantic similarity of the sentence to the query and the title, if any.

Using the abovementioned features, we are able to give a score to each sentence in all documents signifying their importance. The next section describes how the scoring takes place.

### 3.4 Scoring the Sentences:

The score for each sentence (score(i)), is generated based on the linear combination of the weighted features computed as described in the previous steps. The formula used for scoring each sentence is:

$$\text{Score(i)}=\frac{(\alpha Sim(s_i ,\text{T})+\beta Sim(s_i ,Q))\ n(s_i)\ (F_{NE}(s_i)+1)\ P(s_i)}{N(NE+1)}$$

Where:

- N is the total number of sentences in the document
- $n(s_i)$ is the number of sentences that have semantic similarity score bigger than a pre-defined threshold value
- $P(s_i)$ is the sentence position weight. For simplicity.
- $Sim(s_i ,T)$ and $Sim(s_i ,Q)$ are for the Semantic Similarity between the Title and the Query, respectively, and the sentence (i) determined using the sentence-sentence semantic similarity previously described.
- NE is the number of Named Entities in the document
- $F_{NE}(s_i)$ is the number of Named Entities contained in the sentence (i)

The rationale behind the preceding is to allow the score assigned to the sentence si very much dependent on the evaluation of the semantic similarity of si to both the title and the query using a convex combination of both entities. This output is weighted by n(si), which expresses, at some extent, the frequency of the sentences in the document(s) that are semantically similar to si up to some threshold μ, as well as the number of Named Entities in the sentence and its position. The positioning parameter is motivated by the observation that usually, beginning and end of the document contains more information regarding the context of the underlying document (s) as authors attempt to provide concise overview at the beginning and concluding remarks at the end. But, obviously this is very much context dependent. The weighting parameters α and β ($\alpha + \beta = 1$) are left open to the choice of the user depending on his/her prior knowledge about the relevance of the title and/or query. In the absence of any further evidence, the default values are 0.5 each, which is in agreement with the principle of insufficient reason in statistics.

### 3.5 Generating Summaries:

A summary is generated by choosing the most important sentences in a document (or the highest scoring) and arranging them in chronological order to insure the readability of the generated summary. Multi-document summaries are generated in a similar fashion by computing sentences scores in each document separately and then choosing the highest scoring sentences from all documents to generate multi-document summaries.

Handling the information redundancy between sentences and within each sentence was not completed in time and thus was not part of the system we used to participate in TAC 2008.

## 4. Evaluation

To evaluate our system, we participated in TAC 2008 for the first time even though some major components were not fully implemented in our system yet (i.e. redundancy checking). Next, we present results obtained from the automatic evaluation performed by NIST using ROUGE [8] and BE [9] metrics, and the manual responsiveness measure.

### 4.1 Test Data and Metrics:

For the TAC 2008 update task, we adopted the Jiang & Conrath [6] method when computing the semantic similarity between words. The redundancy handling component was not completed in time and thus the system used when participating in TAC 2008 did not handle sentences redundancy.

The provided test dataset comprised 48 topics. Each topic had a topic statement and 20 relevant documents which had been divided equally into 2 sets: A and B. The set A always chronologically precedes the documents in set B. The provided test dataset was taken from the AQUAINT-2 collection of news articles.[2]

All of the submitted summaries were truncated to 100 words. NIST conducted manual evaluation of summary content based on the Pyramid Method. Four different NIST assessors would create 100-word model summaries for each document set that addresses the information need expressed in the topic statement.

Each participant team was requested to submit up to 3 runs ranked by priority (1-3). Our team submitted two runs: one (run # 1) has more weight given to the topic statement, and the other (run # 34) has more weight given to the headlines. In the abovementioned scoring formula, α was given a value of 0.75 for run 1, and 0.25 for run 34.

[2] See http://www.nist.gov/tac/tracks/2008/summarization/update.summ.08.guidelines.html for more details about the Task.

### 4.2 Results:

In the update task of TAC 2008, 57 peer summaries were manually evaluated with the pyramid method, and 71 were evaluated using ROUGE and the Basic Elements evaluation package [9].

Table 1 shows the average Recall, Precession and F-measure for the Rouge1, Rouge2, and RougeSU4 evaluations on the two runs we submitted. It can be noted that in both runs, the system generally ranked higher in Recall than Precession. This suggests that the system is better at finding relevant content than it is at removing irrelevant content. Also, it can be noted that the run which more weight given to the topic statement generally achieved better ROUGE scores than the other run with more weight given to the headlines.

| | **Run 1** | | | **Run 34** | | |
|---|---|---|---|---|---|---|
| **ROUGE** | **Avg R** | **Avg P** | **Avg. F** | **Avg R** | **Avg P** | **Avg F** |
| **1** | 0.34463 | 0.33866 | 0.34148 | 0.34022 | 0.33372 | 0.33680 |
| **2** | 0.08091 | 0.07933 | 0.08008 | 0.08080 | 0.07912 | 0.07991 |
| **SU4** | 0.11852 | 0.11634 | 0.11737 | 0.11706 | 0.11471 | 0.11583 |

**Table1:** The Rouge Scores obtained by our system in the two runs we submitted.

Table 2 shows the automated evaluations average scores obtained by our submitted runs (with their ranks) in comparison with the 71 peer summaries submitted by the rest of the participants.

| **Evaluation** | **Run (1)** | **Run (34)** | **Best** | **Worst** |
|---|---|---|---|---|
| **ROUGE2-R** | 0.08091<br>(25/71) | 0.08080<br>(26/71) | 0.10382 | 0.03343 |
| **ROUGESU4-R** | 0.11858<br>(23/71) | 0.11713<br>(29/71) | 0.13646 | 0.06517 |
| **BE** | 0.04964<br>(24/71) | 0.04903<br>(28/71) | 0.06462 | 0.01337 |

**Table2**: the automated scores (and ranks) obtained by our system in comparison with the rest.

The evaluation in TAC2008 included human judgments of linguistic quality. Table 3 shows the results and the rank of our system in respect with the rest in the manual evaluation. The metrics shown in the table are: responsiveness which is how well the summary addresses the user's information need; and linguistic quality. The linguistic quality score is guided by consideration of the following factors:

1. Grammaticality
2. Non-redundancy
3. Referential clarity
4. Focus
5. Structure and Coherence

with scores between 1 (very poor) and 5 (very good).

| | **Run (1)** | **Run (34)** | **Best** | **Worst** |
|---|---|---|---|---|
| **Avg Linguistic Quality** | 2.719<br>(12/58) | 2.76<br>(11/58) | 3.073 | 1.312 |
| **Overall Responsiveness** | 2.427<br>(15/57) | 2.385<br>(18/57) | 2.667 | 1.198 |

**Table 3** : Manual Evaluation Results

## 5. Future Work

We plan to add and improve many aspects of the system we developed, especially in the post-processing part. Among the ideas we plan to integrate are the following:

- Implement redundancy checking and remove repeated information. We think that implementing this feature will greatly enhance the evaluation results. This should be done from two different perspectives: First, removing repeated or non-essential content from within sentences such as relative clauses (which can be done in the last stage just before choosing the summary highest scoring sentences by adding a new metric: redundancy penalty affecting the repeated sentences score). Second, relating the chosen summary sentences with each other and trying to maximize the information content diversity between sentences to achieve the highest possible comprehensiveness in the generated summary. To achieve the later, the semantic similarity between the summary candidate sentences can be checked against a previously set threshold and thus reducing the score for those sentences containing repeated data.
- Try to find a method to automatically optimize the weight of the dynamic features. Currently, the weights are assigned manually based on the user's observations. Implementing this will require great deal of analysis to the syntax of the text in each sentence, which has not been deeply explored in our system.
- Compressing the summary sentences to allow for more information to be presented in the summary at the same or shorter length. Syntactic trimming which has been studied in previous work [3] is what we are currently exploring and hoping to improve and implement in our system.

- Meeting the word limit in our system was achieved by simply iterating through all the highest scoring sentences to replace the last summary sentence with the next shorter and high scoring sentence. This means that in some cases, none of the sentences are chosen (if replacing the last summary sentence with any other will yield a summary longer than the required word-limit) and thus sentences with valuable and relevant content to the user query are not added because of their length. This will need further investigation and can be partially overcome by generating shorter forms of long sentences (compressing the summary sentences) and eliminating non-important sentences before the processing stage using "shallow parsing " techniques similar to [10].
- We plan to use co reference resolution to enhance the quality of our generated summaries. For example, some sentences might contain references to important entities such as "President Bush" in the form of one word "he". We think that replacing the pronoun with the Named Entities before processing the summaries should give better scores for our summaries [11].

## 6. Conclusion

In this paper we present our on-going work on building a query-focused multi-document summarization system and the evaluation results for the system in the update task of TAC 2008. The results suggest that our overall system rank can be placed in the middle tier when compared with all the participants in the task for this year. In future work, we plan to apply and experiment with more detailed measures to handle different aspects such as redundancy, comprehensiveness and length, and automatic weight optimization for the dynamic features.

## References

1. Radev, D., et al., *MEAD ReDUCs: Michigan at DUC 2003.* Proceedings of DUC 2003, 2003.

2. Radev, D.R. and G. Erkan, *The University of Michigan at duc2004.* Proceedings of Document Understanding Conference Workshop, 2004: p. 120-127.

3. Zajic, D., *Multiple Alternative Sentence Compressions as a Tool for Automatic Summarization Tasks*, in *Department of Computer Science*. 2007, University of Maryland.

4. GATE. *GATE - General Architecture for Text Engineering*. 2007; Available from: www.gate.ac.uk.

5. Sekine, S. and C. Nobata, *Sentence Extraction with Information Extraction Techniques.* Workshop on Text Summarization 2001, 2001.

6. Jiang, J. and D. Conrath, *Semantic Similarity based on Corpus Statistics and Lexical Taxonomy.* Proceedings of International Conference Research on Computational Linguistics,Taiwan, 1997.

7. Lin, D., *An information-theoretic definition of similarity.* Proc. 15th International Conference on Machine Learning, 1998: p. 296-304.

8. Lin, C.-Y., *ROUGE: a Package for Automatic Evaluation of Summaries.* Proceedings of the Workshop on Text Summarization Branches Out (2004), 2004.

9. Hovy, E., C. Lin, and L. Zhou, *Evaluating DUC 2005 using Basic Elements.* Proceedings of the HLT/EMNLP Workshop on Text Summarization DUC 2005, 2005.

10. Dunlavy, D., et al., *Performance of a Three-Stage System for Multi-Document Summarization.* 2003.

11. Conroy, J.M., et al., *Left-Brain/Right-Brain Multi-Document Summarization.* Document Understanding Conference Workshop at HLT/NAACL 2004, 2004.
12. F. Bond, Determiners and Number in English contrasted with Japanese, as exemplified in Machine Translation, University of Queensland, 2001